Theoretical Computer Science 582 (2015) 60–69

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Core words and Parikh matrices

Wen Chean Teh [a,*], Kiam Heong Kwa [b]

[a] School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Malaysia
[b] Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 25 April 2014
Received in revised form 12 December 2014
Accepted 21 March 2015
Available online 31 March 2015
Communicated by M. Ito

Keywords:
Parikh matrix
Subword
Ambiguity
Core of a word

ABSTRACT

Parikh matrices have been widely investigated due to their applicability in arithmetizing words by numbers. This paper introduces the core of a binary word, which captures the essential part of a word from the perspective of its Parikh matrix. Additionally, the stronger notion of core M-unambiguity is introduced and the characterization of core M-unambiguous binary words is obtained. Finally, a generalization of the core of a binary word and some of its interesting properties are investigated.

1. Introduction

The optimal number of subwords of a word required to determine the word is still an open question [12]. The Parikh matrix of a word, which contains the numbers of scattered occurrences of some words of the word, was initially introduced in [16] to deal with questions of this nature. Since then there have been extensive investigations on various questions concerning this class of matrices and its variants [6,10,11,15,20,21,23,24]. Generally, a Parikh matrix does not determine a word. Hence one of the most studied questions in this direction is the injectivity problem of the Parikh matrix mapping and the related M-equivalent classes of words, each of which consists of words having a common Parikh matrix. A word is M-unambiguous if and only if its M-equivalence class is a singleton. The injectivity problem and the characterization of M-unambiguity have received great interest [4,5,7,9,14,19,22,25–27] but proved to be elusive. Despite much effort, not much is known about M-equivalent words other than for binary words [3,7,8,13].

This paper aims to contribute to the injectivity problem by introducing the core of a binary word in Section 3. We will see that the injectivity problem for binary words reduces to the injectivity problem restricted to these binary core words. Using the binary core words, each M-equivalent class of a binary word can be identified with a certain set of partitions of a natural number. Furthermore, a characterization of the set of binary words that are *core M-unambiguous* is obtained.

In Section 4, we investigate the notion of the core of a word relative to a given word for *any* alphabet, of which the core of a binary word is a special case. Section 5 gives some sufficient or necessary conditions for 1-equivalence for higher alphabets in terms of our relativized cores of words. We conclude in Section 6 mentioning our future direction. For completeness, the next section reviews the formal definitions of a Parikh matrix, M-equivalence and M-ambiguity and states some of their elementary properties.

* Corresponding author.
E-mail addresses: dasmenteh@usm.my (W.C. Teh), khkwa@um.edu.my (K.H. Kwa).

http://dx.doi.org/10.1016/j.tcs.2015.03.037
0304-3975/

2. Subwords and Parikh matrices

The reader is referred to [18] for language theoretic notions and results not detailed here.

Suppose Σ is an alphabet. The set of words over Σ is denoted Σ^*. The empty word is denoted λ. Let Σ^+ denote the set $\Sigma^*\backslash\{\lambda\}$. If $v, w \in \Sigma^*$, the concatenation of v and w is denoted by $v \cdot w$, or simply vw. An *ordered alphabet* is an alphabet $\Sigma = \{a_1, a_2, \dots, a_s\}$ with a total ordering on it. For example, if $a_1 < a_2 < \dots < a_s$, then we may write $\Sigma = \{a_1 < a_2 < \dots < a_s\}$. For $1 \le i \le j \le s$, let $a_{i,j}$ denote the word $a_i a_{i+1} \dots a_j$. If $w \in \Sigma^*$, then $|w|$ is the length of w and $w[i]$ is the $(i+1)$-th letter of w. Suppose $\Gamma \subseteq \Sigma$. The projective morphism $\pi_\Gamma \colon \Sigma^* \to \Gamma^*$ is defined by

$$\pi_\Gamma(a) = \begin{cases} a, & \text{if } a \in \Gamma \\ \lambda, & \text{otherwise.} \end{cases}$$

We may write $\pi_{a,b}$ for $\pi_{\{a,b\}}$.

Definition 2.1. A word w' is a *subword* of a word $w \in \Sigma^*$ iff there exist words $x_1, x_2, \dots, x_n, y_0, y_1, \dots, y_n \in \Sigma^*$, some of them possibly empty, such that

$$w' = x_1 x_2 \dots x_n \text{ and } w = y_0 x_1 y_1 \dots y_{n-1} x_n y_n.$$

In the literature, our subwords are usually called "scattered subwords". The number of occurrences of a word u as a subword of w is denoted by $|w|_u$. As a special case, $|w|_\lambda = 1$ for all $w \in \Sigma^*$. Note that two occurrences of u are considered different if they differ by at least one position of some letters. The *support* of w, denoted $\text{supp}(w)$, is the set $\{a \in \Sigma \mid |w|_a \neq 0\}$.

Suppose $\Sigma = \{a_1 < a_2 < \dots < a_s\}$ is an ordered alphabet. The *Parikh mapping* $\Psi \colon \Sigma^* \to \mathbb{N}^s$ is defined by

$$\Psi(w) = (|w|_{a_1}, |w|_{a_2}, \dots, |w|_{a_s}).$$

We say that $(|w|_{a_1}, |w|_{a_2}, \dots, |w|_{a_s})$ is the Parikh vector of w. The classical Parikh's Theorem [17] states that the image of a context free language under the Parikh mapping is a *semilinear* set.

For any integer $k \ge 2$, let $\mathcal{M}_k$ denote the multiplicative monoid of $k \times k$ upper triangular matrices with nonnegative integral entries and unit diagonal. The following Parikh matrix mapping is a generalization of the Parikh mapping, where the Parikh vector of a word is contained in the second diagonal of the Parikh matrix of the word.

Definition 2.2. (See [16].) Suppose $\Sigma = \{a_1 < a_2 < \dots < a_s\}$ is an ordered alphabet. The *Parikh matrix mapping*, denoted by Ψ_Σ, is the morphism

$$\Psi_\Sigma \colon \Sigma^* \to \mathcal{M}_{s+1}$$

defined as follows:
if $\Psi_\Sigma(a_q) = (m_{i,j})_{1 \le i,j \le s+1}$, then $m_{i,i} = 1$ for each $1 \le i \le s+1$, $m_{q,q+1} = 1$ and all other entries of the matrix $\Psi_\Sigma(a_q)$ are zero.

A matrix $N \in \mathcal{M}_{s+1}$ is a *Parikh matrix* iff $N = \Psi_\Sigma(w)$ for some $w \in \Sigma^*$.

Theorem 2.3. *(See [16].) Suppose $\Sigma = \{a_1 < a_2 < \dots < a_s\}$ is an ordered alphabet and $w \in \Sigma^*$. The matrix $\Psi_\Sigma(w) = (m_{i,j})_{1 \le i,j \le s+1}$ has the following properties:*

- *$m_{i,i} = 1$ for each $1 \le i \le s+1$;*
- *$m_{i,j} = 0$ for each $1 \le j < i \le s+1$;*
- *$m_{i,j+1} = |w|_{a_{i,j}}$ for each $1 \le i \le j \le s$.*

Example 2.4. Suppose $\Sigma = \{a < b < c\}$ and $w = ababcc$. Then

$$\begin{aligned}
\Psi_\Sigma(w) &= \Psi_\Sigma(a)\Psi_\Sigma(b)\Psi_\Sigma(a)\Psi_\Sigma(b)\Psi_\Sigma(c)\Psi_\Sigma(c) \\
&= \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 2 & 3 & 6 \\ 0 & 1 & 2 & 4 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.
\end{aligned}$$